# Mardi 24 septembre 2024

Hier nous avons terminé la journée sur une réflexion concernant le développement et le monitoring du copilote. Initialement nous pensions que :

* Il pourrait y avoir un risque lié à l’attribution des droits d’accès aux données du SharePoint au sein du Copilote et engendrer des fuites de données.
  + Or, d’après la documentation le copilote utiliserait une authentification via Microsoft Entra ID, qui vérifie si l’utilisateurs aux droits nécessaires pour obtenir une réponse.
* La quantité de données présentent sur le SharePoint pourrait influencer la précision du copilote.
  + Or, selon Microsoft, le facteur qui relierait les données à la précision serait en réalité la correcte labellisation des données au sein du SharePoint.
* Le Prompt engineering devrait être amélioré (grâce à une boucle d’amélioration automatique), afin que les collaborateurs qui manqueraient de compétence en prompt engineering puissent quand même profiter du copilote au maximum de ses capacités.
  + Or, le copilote est conçu pour améliorer automatiquement les prompts basés sur les interactions avec les utilisateurs, incluent une optimisation des requêtes via l’API Graph. Rajouter une boucle d’amélioration pourrait être redondant et risque de nuire à la performance si elle interfère avec le système d’amélioration déjà intégré.
* La technicité d’un document influait sur la capacité de retrieval du copilote.
  + Or, le facteur qui influe la capacité de retrieval du copilote serait en réalité la structure du document analysé.

En creusant plus profondément dans la documentation, j’ai trouvé une information cruciale :

Alors que je cherchais à me documenter concernant la capacité de retrieval du copilote (sur le GPT sur la documentation Microsoft) il est mentionné que les fichiers dépassant les 3 Mo ne sont pas pris en compte par les réponses génératives.

* Serait-il possible d’ajouter une boucle afin de désactiver les réponses génératives ?

Nous avons eu un point avec Jérôme afin de discuter de mon avancée sur le sujet Copilote. Je lui ai fais part de ma méthodologie d’utiliser les GPTs sur OpenAI avec pour source de savoir les documentations de Microsoft Copilote Studio ainsi que Power Automate.

L’importance de l’information concernant la capacité du copilote à n’utiliser que les fichiers d’une taille maximum de 3Mo pour les réponses génératives nous a permis de soulever différents points :

* Premièrement s’il était possible de contourner cette problématique de blockage pour les fichiers dépassant les 3Mo.
* Ensuite sur l’architecture de la vectorisation utilisé au sein du Copilote.
  + Réflexion : il semblerait que Microsoft ait conceptualisé une architecture de vectorisation différente de celle plus largement connue (Langchain). Nous supposons que l’API Graph Navigue, peut-être grâce à une vectorisation, à travers le sharepoint afin de trouver précisément le fichier nécessaire pour répondre à la requête de l’utilisateur. Cela évite de se retrouver avec une vectore base énorme (vectorisation de chunk de +300 000 documents) qui apporterait un taux d’hallucination rendant le copilote inexploitable.

Il semblerait que l’indexation sémantique soit la solution adaptée à notre problème, sinon nous pouvons aussi indexer nos données sur Azure Copilot.